AI RELIABILITY MONITORING

Guardian

Know when your AI systems are underperforming, deceiving, or drifting. Real-time monitoring with research-backed detection methods.

AI systems fail silently

Production AI has failure modes that traditional monitoring can't detect. Models sandbag to avoid scrutiny. They hallucinate confidently. They drift as providers update them. By the time you notice, the damage is done.

96%	<50ms	24/7
Sandbagging detection accuracy	Detection latency overhead	Continuous monitoring

Sandbagging

Detect when models deliberately hide capabilities or underperform

Hallucination

Track confidence calibration and factual accuracy in real-time

Drift

Catch behavioral changes from provider updates or distribution shift

Compliance

Audit-ready reports for regulators and stakeholders

BUILT ON

rotalabs-probe open source toolkit. Peer-reviewed research. Verify our methods.

INTEGRATES WITH

Any model provider—OpenAI, Anthropic, open source, or your fine-tuned models

01 — DETECTION CAPABILITIES

Four failure modes. Continuous detection.

Traditional monitoring tracks latency and errors. Guardian detects the silent failures that destroy trust in AI systems.

S

Sandbagging Detection

Models sometimes deliberately hide capabilities or underperform on evaluation tasks while performing normally otherwise. Guardian uses metacognitive probing techniques to detect this deceptive behavior with 96% accuracy. Based on peer-reviewed research from Rotalabs.

H

Hallucination Monitoring

Track confidence calibration across all outputs. Detect when models express high confidence in incorrect information. Alert on sudden increases in hallucination rates. Distinguish between uncertainty and confident confabulation.

D

Drift Detection

Automatically establish behavioral baselines during initial deployment. Detect when model behavior deviates—from provider updates, distribution shift, or prompt injection. Track performance degradation before it impacts users.

C

Compliance Reporting

Generate audit-ready reports for regulators. Document model behavior, decision rationale, and reliability metrics over time. Meet EU AI Act, OCC SR 11-7, and MAS FEAT requirements for AI transparency.

Research Foundation: Rotalabs

Guardian's detection methods are built on peer-reviewed research from Rotalabs, our open source AI safety research division. The sandbagging detection techniques have been validated against frontier models and published for community verification.

Open source: rotalabs-probe toolkit available at rotalabs.ai. Verify our methods. Contribute improvements.

02 — PLATFORM INTEGRATION

Guardian in the Trust Intelligence Platform

Guardian doesn't work in isolation. It's the reliability layer for the entire AI operations stack.

AgentOps — Agent Registry & Policy → **Guardian** — Continuous Monitoring → **Steer** — Runtime Correction → **Flight Recorder** — Audit Trail

Connected to AgentOps

Every agent registered in AgentOps is automatically monitored by Guardian. Policy violations trigger alerts. Agent health scores feed into governance dashboards.

Feeds Steer

When Guardian detects anomalies, Steer can automatically adjust agent behavior—increase guardrails, route to human oversight, or activate fallback modes.

Writes to Flight Recorder

All detection events, confidence scores, and behavioral metrics are captured in the Agent Flight Recorder for audit and forensic analysis.

Informs Trust Cascade

Agent reliability scores influence Trust Cascade routing. Agents with declining scores get routed to higher verification levels automatically.

03 — INDUSTRY USE CASES

Where Guardian delivers value

FINANCIAL SERVICES

Trading System Monitoring

Detect when AI trading models drift from expected behavior. Alert before anomalous trades execute. Meet MRM requirements for continuous model validation.

HEALTHCARE

Clinical Decision Support

Monitor diagnostic AI for hallucination in recommendations. Track confidence calibration on medical advice. Alert clinicians when AI reliability drops.

INSURANCE

Underwriting AI

Detect bias drift in underwriting models. Monitor for sandbagging on compliance evaluations. Document decision rationale for regulatory audits.

CUSTOMER SERVICE

Chatbot Reliability

Track hallucination rates in customer-facing AI. Detect when chatbots give confident wrong answers. Maintain customer trust through reliability monitoring.

04 — TECHNICAL SPECIFICATIONS

Enterprise-ready architecture

CAPABILITY	SPECIFICATION
Detection Latency	<50ms overhead on inference calls
Supported Models	OpenAI, Anthropic, Cohere, open source (Llama, Mistral), custom fine-tuned
Integration	REST API, Python SDK, JavaScript SDK, OpenTelemetry
Deployment	SaaS, private cloud (AWS, GCP, Azure), on-premise, air-gapped
Alert Channels	Slack, PagerDuty, email, webhooks, SIEM integration
Data Retention	Configurable: 30 days to 7 years (compliance requirements)
Compliance	SOC 2 Type II, GDPR, HIPAA-ready, EU AI Act compatible

05 — PRICING

Plans for every scale

Starter

$500/mo

1 model, 100K inferences/mo, dashboard, basic alerts. For teams getting started with AI reliability.

Pro

$2,000/mo

5 models, 1M inferences/mo, advanced alerts, API access, compliance reports. For production AI.

Enterprise

Custom

Unlimited models, on-premise deployment, SSO, SLA, dedicated support. For regulated industries.

Pricing is indicative. Contact us for volume discounts and custom requirements.

See Guardian in action

Live demo with simulated anomalies on your use case.

Request Demo

Learn More